Summary of issues

1. (2.02). What's "m"?

Basics of Information Theory (5)

❑ From the axioms we can derive the following *properties* (assuming independent events, same *p*):

1. $I(p^2) = I(p \cdot p) = I(p) + I(p) = 2 \cdot I(p)$
2. Thus in general, $I(p^n) = n \cdot I(p)$
3. $I(p) = I\left(\left(p^{1/m}\right)^m\right) = mI\left(p^{1/m}\right) \Rightarrow I\left(p^{1/m}\right) = \frac{1}{m} I(p)$ m?
4. Thus in general, $I\left(p^{n/m}\right) = \frac{n}{m} \cdot I(p)$
5. By continuity, for $0<p\leq 1$, and $a>0$ real number

$$I(p^a) = a \cdot I(p)$$

2. (2.02). How to understand this red mark area? What is the definition of "entropy"?

Entropy Theory (7)

❑ With the above definitions we have that

$$S(P) =< I(P) >$$

❑ In other words, *the entropy of a probability distribution is the expected value of the information of the distribution*

The bridge between information theoretical and physical entropies lies in *how hard it is to describe a physical state or process*. The amount of information it takes to describe something is proportional to its entropy.

23

3. (2.02). Can u help me to clear the prove?

Maximum entropy

$$S(P) - \log(n) = \sum_{i=1}^{n} p_i \log(1/p_i) - \log(n) =$$

$$= \sum_{i=1}^{n} p_i \log(1/p_i) - \log(n) \sum_{i=1}^{n} p_i =$$

$$= \sum_{i=1}^{n} p_i \log(1/p_i) - \sum_{i=1}^{n} p_i \log(n) =$$

$$= \sum_{i=1}^{n} p_i \left[\log(1/p_i) - \log(n)\right] =$$

$$= \sum_{i=1}^{n} p_i \left[-\log(p_i) + \log(1/n)\right] =$$

$$= \sum_{i=1}^{n} p_i \log\left(\frac{1/n}{p_i}\right) \underset{Gibbs}{\leq} 0$$

Equality holds only when $p_i=1/n$ for all i

4. **(2.03). Is "compress" means get everything in order, and entropy will be the most messy value?**

Encoding efficiency vs. entropy

- In building encoding schemes, we have to use our best understandings of the *structure* of a data stream (in other words, we want to use our best *probability model* of the data stream)

- The *entropy* gives us a lower bound on our encoding efficiency. Thus, if we want to improve our schemes, we will have to develop successively better probability models

5. **(2.04). Why "S=2 bits" ?**

Entropy and coding – an example

- Stochastic process that generates random sequences of symbols a, b, c, d
- Suppose – to start with – that it is unknown with which probabilities symbols are generated
- Then the best guess is to assign probability 1/4 to each event, and that subsequent symbols are independent
- Then information gained in any possible observation of a single symbol is $\log_2(4) = 2$ bits. Therefore, the entropy S = 2 bits. ?
- Reasonable, as we can simply code our four symbols with binary codewords 00, 01, 10, 11

6. **(2.04). Can not prove this, can u help me to go through the prove?**

Applying the MEP – Economics

- Putting in the constraints, we have

$$1 = \sum_i p_i = \sum_i e^{-\lambda_0} e^{-\lambda i} = e^{-\lambda_0} \sum_i e^{-\lambda i}$$
$$\frac{M}{N} = \sum_i p_i \cdot i = \sum_i e^{-\lambda_0} e^{-\lambda i} \cdot i = e^{-\lambda_0} \sum_i e^{-\lambda i} \cdot i \qquad (5)$$

- For large M, we can approximate as

$$\sum_{i=0}^{M} e^{-\lambda i} \approx \int_0^M e^{-\lambda x} dx \approx \frac{1}{\lambda}$$
$$\sum_{i=0}^{M} e^{-\lambda i} \cdot i \approx \int_0^M x e^{-\lambda x} dx \approx \frac{1}{\lambda^2} \qquad (6)$$

Check this.

7. (2.05). Can not prove this.

Mutual information

See Lecture 2.03

$$I[X_1;X_2] = S[X_1] + S[X_2] - S[X_1,X_2] =$$
$$= \sum_{x_1} p(x_1)\log\frac{1}{p(x_1)} + \sum_{x_2} p(x_2)\log\frac{1}{p(x_2)} - \sum_{x_1x_2} p(x_1x_2)\log\frac{1}{p(x_1x_2)} =$$
$$= \sum_{x_1x_2} p(x_1x_2)\log\frac{p(x_1x_2)}{p(x_1)p(x_2)} = K[P(X_1)P(X_2);P(X_1X_2)] \geq 0$$

Prove this

- The *mutual information* is the information we get when we replace the separate distributions $P(X_1)$ and $P(X_2)$ as the description of the system, with the correct joint distribution $P(X_1,X_2)$

8. (2.05). Can not prove this.

Correlation information

- *Correlation information from length n*

$$k_n = \sum_{x_1\ldots x_{n-1}} p(x_1\ldots x_{n-1})K[P^{(0)};P] =$$
$$= \sum_{x_1\ldots x_{n-1}} p(x_1\ldots x_{n-1})\sum_{x_n}\frac{p(x_1\ldots x_{n-1}x_n)}{p(x_1\ldots x_{n-1})}\log\frac{p(x_1\ldots x_{n-1}x_n)p(x_2\ldots x_{n-1})}{p(x_1\ldots x_{n-1})p(x_2\ldots x_{n-1}x_n)}$$

$$k_n = -S_n + 2S_{n-1} - S_{n-2} = -\Delta S_n + \Delta S_{n-1} = -\Delta^2 S_n \quad (n = 2,3,\ldots)$$

See Fig. 1

Prove this

- Correlation information from length 1

$$k_1 = K[P_1^{(0)};P_1] = \sum_{x_1} p(x_1)\log\frac{p(x_1)}{1/\nu} = \log\nu - S_1$$

9. (2.06). How to calculate "I"? Is red area wrong with expression?

Markov processes and entropy

- One finds that p(a) = p(b) = p(c) = 1/3

- There is an element of surprise only when the process is in state a or b, so the entropy turns out to be

$$s = p(a)\,I_a + p(b)\,I_b + p(c)\,I_c = \frac{1}{3}\cdot 1 + \frac{1}{3}\cdot 1 + \frac{1}{3}\cdot 0 = \frac{2}{3} \text{ bits}$$

10. (2.06). Sorry to my poor math calculations, but how did this one get?

Crystal

- Using eqs.

$$K[P^{(0)};P] = \sum_{x_n} p(x_n \mid x_1 x_2 \ldots x_{n-1}) \log \frac{p(x_n \mid x_1 x_2 \ldots x_{n-1})}{p(x_n \mid x_2 \ldots x_{n-1})}$$

$$k_n = \sum_{x_1 \ldots x_{n-1}} p(x_1 \ldots x_{n-1}) K[P^{(0)};P]$$

- Plugging eq. (2) in the above equation, we can compute the correlation information from length 2 :

$$k_2 = \sum_{x_1} p(x_1) \sum_{x_2} p(x_2 \mid x_1) \log \frac{p(x_2 \mid x_1)}{p(x_2)} = \log 2 = 1 \text{ (bit)}$$

???

11. (2.08) How can I get the table below?

Elementary cellular automata

- The simplest class of CA is based on a 1D lattice with 2 states per cell and a nearest-neighbour rule for the dynamics
- We call this class of CA, Elementary Cellular Automata (ECA)
- Such a rule is fully determined by specifying the next state of a cell for each of the 8 possible local states that describes the present state of the local neighbourhood
- An example of such a specification is shown in the following Table

t	111	110	101	100	011	010	001	000
$t+1$	0	1	1	0	1	1	1	0

12. (2.08) For Lecture 2.08 could you remind me about four Class of Rules and what is the main differences between them?

13. About the first tutorial, can you help me to clear the logic of question 2? And what the meaning by "m" and "v"?

My understanding will be: **"m"** is the length of sequence that must occur according to the model (shortest) and for **"v"** is the alphabet that can be generated in this model. I'm not sure is this correct.

14. (4.01) How to understand "Clustering Coefficient"?

15. (4.01) As my understanding here should be (N(N-1)/2)-n, is this correct?

Random Graphs: Average number of edges ($\langle n \rangle$)

Due to p, the total number of edges is a random variable. The average number of edges is,

$$\langle n \rangle = p\,(N(N-1)/2)\,.$$

The probability of creating a graph G_0 with N nodes and n edges is,

$$P(G_0) = p^n(1-p)^{N(N-1)/2-n},$$

where p is the probability of an edge being created and $(1-p)$ is the probability of it not being created.